# Detection of Tumor Morphology Mentions in Clinical Reports in Spanish Using Transformers

Guillermo López-García[1]*, José M. Jerez[1], Nuria Ribelles[2], Emilio Alba[2], and Francisco J. Veredas[1]

[1] Departamento de Lenguajes y Ciencias de la Computación, Universidad de Málaga, 29071 Málaga, Spain

[2] Unidad de Gestión Clínica Intercentros de Oncología, Instituto de Investigación Biomédica de Málaga (IBIMA), Hospitales Universitarios Regional y Virgen de la Victoria, 29010 Málaga, Spain

**Abstract.** The aim of this study is to systematically examine the performance of transformer-based models for the detection of tumor morphology mentions in clinical documents in Spanish. For this purpose, we analyzed 3 transformer models supporting the Spanish language, namely multilingual BERT, BETO and XLM-RoBERTa. By means of a transfer-learning-based approach, the models were first pretrained on a collection of real-world oncology clinical cases with the goal of adapting transformers to the distinctive features of the Spanish oncology domain. The resulting models were further fine-tuned on the Cantemist-NER task, addressing the detection of tumor morphology mentions as a multi-class sequence-labeling problem. To evaluate the effectiveness of the proposed approach, we compared the obtained results by the domain-specific version of the examined transformers with the performance achieved by the general-domain version of the models. The results obtained in this paper empirically demonstrated that, for every analyzed transformer, the clinical version outperformed the corresponding general-domain model on the detection of tumor morphology mentions in clinical case reports in Spanish. Additionally, the combination of the transfer-learning-based approach with an ensemble strategy exploiting the predictive capabilities of the distinct transformer architectures yielded the best obtained results, achieving a precision value of 0.893, a recall of 0.887 and an F1-score of 0.89, which remarkably surpassed the prior state-of-the-art performance for the Cantemist-NER task.

**Keywords:** Transformers · Tumor morphology mentions · Natural language processing · Deep learning · Oncology

## 1 Introduction

There is a significant demand for the automated analysis of the information stored in electronic health records (EHRs) to improve patient care. EHRs contain heterogeneous data whose volume is consistently growing, including *free-text* documents that, using domain-specific vocabulary and terminology, store

* Corresponding author (guilopgar@uma.es).

crucial patient information about clinical examinations, radiology reports, discharge summaries, etc. [2] However, the unstructured nature of the texts makes it particularly challenging to directly extract the relevant medical information these documents contain. In this way, there is a pressing need to automatically transform unstructured clinical text into structured information, which can subsequently serve as support in clinical decision-making and in optimizing the administrative management of the resources of healthcare services, improving many aspects of clinical care [3].

According to the World Health Organization (WHO), cancer is a leading cause of mortality worldwide, producing around 10 million deaths in 2020 [23]. Diagnosis of cancer heavily relies on the pathological examination of tumor samples obtained from biopsies. The resulting observations made by physicians are mainly reported in pathology reports, which correspond to clinical free-text documents stored in EHRs [17]. With the widespread adoption of EHRs as an essential element in oncology information systems, automatically extracting the information contained in cancer-related EHR documents would not only facilitate pathologists daily clinical practice, but also would permit large-scale analysis of the relations between a concrete tumor case and its prognosis, its response to specific treatments, and many other medical aspects [15].

Traditionally, natural language processing (NLP) techniques have been applied to clinical notes with the aim of extracting relevant medical information from free-text documents [8, 16]. More specifically, these techniques have also been adapted to process oncological textual data, contributing to obtain structured representations of the information stored in cancer-related documents [18, 29]. However, the majority of the previous works focus exclusively on medical texts written in English, owing to the limited availability of annotated corpora and additional clinical linguistic resources written in non-English languages, such as Spanish. With nearly 489 million native speakers, Spanish is the second most spoken language in the world in terms of number of native speakers [25]. Given the enormous amount of clinical texts produced in hospitals from Spanish-speaking countries around the globe, there is a considerable interest both in industry and academia to boost the application of NLP technologies to medical documents in Spanish.

With the aim of overcoming this issue, last year the *CANcer TExt MIning Shared Task* (CANTEMIST) was carried out [15], constituting the first shared task specifically focused on the development of automatic systems for extracting relevant clinical information from oncology texts in Spanish. In particular, Cantemist explored the named entity recognition (NER) of tumor morphology mentions in oncology documents in Spanish. The organizers publicly released the Cantemist corpus, a collection of 1301 oncological clinical case reports manually annotated with mentions of tumor morphology. Additionally, the tumor morphology mentions were mapped to a standardized coding vocabulary, specifically the CIE-O—which is the Spanish equivalent of the ICD-O (International Classification of Diseases for Oncology). Within the Cantemist track, three different shared subtasks were proposed: Cantemist-NER, Cantemist-NORM and

Cantemist-CODING. Thus, given a free-text oncology document, the Cantemist-NER task consisted in automatically detecting the tumor morphology mentions contained in the text, whereas the Cantemist-NORM task additionally required assigning the corresponding CIE-O codes to the identified mentions. For its part, the Cantemist-CODING task consisted in assigning a ranked list of CIE-O codes to each text in the Cantemist corpus.

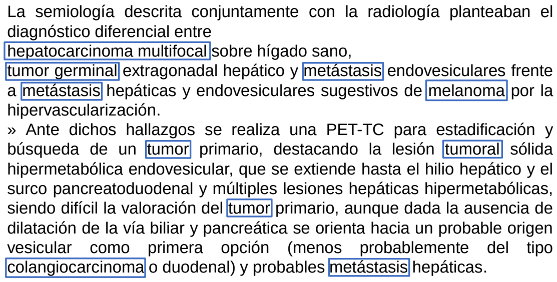
In this work, we have tackled the problem of automatically detecting tumor morphology mentions in oncology cases written in Spanish. For this purpose, we adapted several transformer-based models to the distinctive features of the Spanish oncology domain. By means of a transfer-learning (TL) approach, the models were firstly pretrained on a private collection of real-world oncology clinical cases written in Spanish. The resulting models were further fine-tuned on the Cantemist-NER subtask, addressing the problem as a multi-class sequence-labeling task. Although previous preliminary works have applied BERT-based models to the problem of identifying tumor morphology mentions in clinical documents in Spanish [7, 27], to the best of our knowledge this is the first study that systematically analyzes the performance of different transformer-based architectures for the problem of tumor morphology mentions detection using medical texts in Spanish. Following the proposed TL-based strategy, the transformers analyzed in this work achieved new state-of-the-art (SOTA) performance on the Cantemist-NER subtask. For reproducibility purposes, all the code needed to replicate our work is publicly available at https://github.com/guilopgar/TumorMorphNER.

## 2 Materials and Methods

### 2.1 Corpora

**Galén oncology corpus.** We further pretrained the transformer-based models analyzed in this study using a private corpus of de-identified oncology documents in Spanish retrieved from the Galén Oncology Information System [20]. The corpus corresponds to a compilation of 30.9K real-world clinical cases written by oncologists from the *Hospital Regional Universitario* and the *Hospital Universitario Virgen de la Victoria* in Málaga, Spain, comprising a total of 64.4M words and 437.6M characters.

**Tumor morphology mentions corpus.** We used the Cantemist-NER corpus to fine-tune the models on a tumor morphology mentions detection task. The corpus comprises 1301 oncological cases written in Spanish, which were manually annotated by clinical experts with mentions of tumor morphology [15]. The collection of documents was split into three subsets: the training set, which contains 501 documents and 6396 tumor morphology annotations, the development set, comprising 500 clinical cases and 6001 annotations, and the test set, containing 300 documents and 3633 annotations. The annotations were distributed in BRAT standoff format [22]. Hence, for each annotated tumor morphology, its mention string, its start character offset and its end character offset were provided (see Fig. 1).



La semiología descrita conjuntamente con la radiología planteaban el diagnóstico diferencial entre hepatocarcinoma multifocal sobre hígado sano, tumor germinal extragonadal hepático y metástasis endovesiculares frente a metástasis hepáticas y endovesiculares sugestivos de melanoma por la hipervascularización.

» Ante dichos hallazgos se realiza una PET-TC para estadificación y búsqueda de un tumor primario, destacando la lesión tumoral sólida hipermetabólica endovesicular, que se extiende hasta el hilio hepático y el surco pancreatoduodenal y múltiples lesiones hepáticas hipermetabólicas, siendo difícil la valoración del tumor primario, aunque dada la ausencia de dilatación de la vía biliar y pancreática se orienta hacia un probable origen vesicular como primera opción (menos probablemente del tipo colangiocarcinoma o duodenal) y probables metástasis hepáticas.

**Fig. 1.** Illustration of the tumor morphology annotations from the Cantemist-NER corpus distributed in BRAT format [22], using the *cc_onco93* clinical document from the Cantemist-NER development subset.

### 2.2 Transformer-based models

In the last years, *contextual embeddings* have emerged as a new family of models capable of creating a numerical representation of a word by considering the particular context where the word occurs within the text. Among these new context-aware language models, the Transformer [24] has undoubtedly stood out as the new deep learning SOTA architecture in the field of NLP. BERT [6], RoBERTa [13] and XLM-R [5] are examples of transformer-based models that have become the new SOTA for question answering, text summarization or NER tasks, also in the field of clinical NLP [21, 1, 28]. One of the main characteristics of the Transformer architecture is the self-attention mechanism it uses, which allows the model to parallelize a large part of the network architecture, increasing computing efficiency. Additionally, another distinctive feature of transformer-based models is that they can be pretrained on a general domain corpus and further fine-tuned on a domain-specific corpus to resolve a particular NLP task, following a TL approach.

In this study, we have systematically analyzed the performance of transformers on the tumor morphology mentions detection problem in oncology documents in Spanish. For this purpose, we have examined 3 transformer-based models that support the Spanish language, namely mBERT [6], BETO [4] and XLM-R [5]. To the best of our knowledge, the previous 3 models are the only publicly available transformers including Spanish among their supported languages.

- ***mBERT***: this multilingual transformer uses the same architecture as the BERT-Base model [6], employing a multilingual WordPiece [9] vocabulary of ∼110K subwords. The total number of trainable parameters of the model is ∼177M.

- ***BETO***: the Spanish-BERT model uses a similar architecture to the BERT-Base model [4]. This transformer uses a Spanish vocabulary of ∼31K subtokens, and the total number of trainable parameters is ∼110M.
- ***XLM-R***: the multilingual version of the RoBERTa-Base model [13] was pretrained following a modified version of the XLM approach [12], using a large multilingual SentencePiece [10] vocabulary of ∼250K subtokens, and the total number of trainable parameters is ∼278M.

### 2.3 CRF

Conditional Random Fields (CRF) [11] have been extensively used for sequence-labeling task, such as the tumor morphology mentions detection problem tackled in this study. In this paper, we have used a feature-based CRF model as a competitive baseline with the aim of comparing the performance of transformer-based models with the results obtained by a standard machine learning (ML) method on the Cantemist-NER task.

### 2.4 Transfer-learning approach for automatic tumor morphology mentions detection

In this study, we have applied a TL approach to perform the automatic detection of tumor morphology mentions in Spanish using transformers. Our TL-based strategy consists of two consecutive phases: firstly, the domain-specific pretraining of the transformer-based models, and then the subsequent supervised fine-tuning of the resulting models. In the next paragraphs, both phases are described.

**Unsupervised pretraining.** The 3 transformer-based models examined in this work were further pretrained on a collection of unlabeled real-world oncology clinical cases. Specifically, the two BERT-based models, namely mBERT and BETO, were pretrained on the basis of the Next Sentence Prediction (NSP) task and the Masked Language Model (MLM) objective with the Whole-Word Masking (WWM) modification [6]. On the other hand, the XLM-R model was optimized using the MLM objective with the dynamic masking modification [5].

**Supervised fine-tuning.** In this study, we tackled the automatic detection of tumor morphology mentions in Spanish using transformers. In this way, we addressed this supervised learning task as a multi-class sequence-labeling problem, using the IOB2 [19] tagging scheme. Since the tumor morphology annotations from the Cantemist-NER corpus were distributed in BRAT standoff format (see Fig. 1), we firstly converted them into a different format compatible with the IOB2 scheme. Thus, for each word in a document from the Cantemist-NER corpus, we assigned the label "B" ("Beginning") if it corresponded to the first word of a tumor morphology mention, the label "I" ("Inside") if the word was inside an annotated mention, or the label "O" ("Outside") if the word was not part of any mention. However, transformer-based models do not operate at word-level. Instead, they further break down words into a sequence of subtokens, each model using a specific tokenizer, e.g. XLM-R utilizes a SentecePiece tokenizer with a vocabulary containing ∼250K subwords. In order to effectively leverage the predictive capabilities of transformers when applied to the Cantemist-NER task, we have developed a five-phases approach that performs the supervised fine-tuning of the transformer-based models using a sequence of subtokens annotated with IOB2 labels as input to the models. Fig. 2 shows a visual description of the developed strategy, and each of its five stages is described in the next paragraphs.

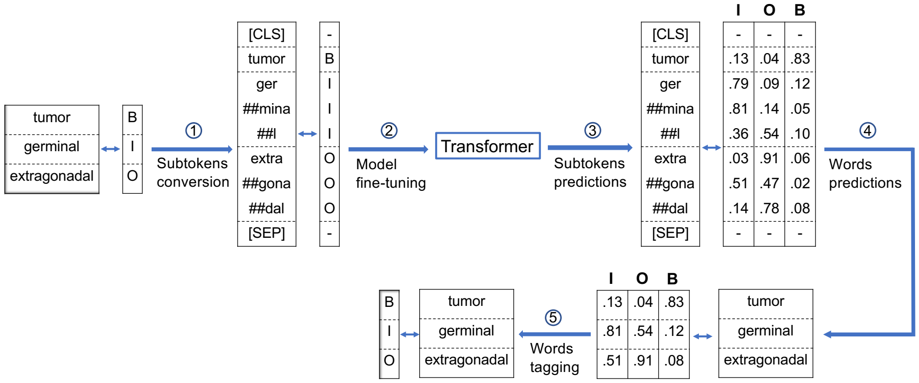


**Fig. 2.** Workflow of the five-phases strategy developed to both fine-tune and evaluate the performance of the transformer-based models on the Cantemist-NER task. For illustration purposes, we used a 3-words fragment of text extracted from the *cc_onco93* clinical case from the Cantemist-NER development corpus (see Fig. 1) as input to the model. The WordPiece tokenizer of the mBERT model was used to generate the subtoken sequence from the input sequence of words. Additionally, the tokenizer added two special tokens (`[CLS]` and `[SEP]`) at the first and last positions, respectively, of the subwords sequence, which are further ignored by the output layer of the model at the time of prediction.

1. **Subtoken-level annotations.** As it was previously specified, transformers further segment words into a sequence of subtokens. For this reason, we converted the IOB2 word-level annotations to subtoken-level. Thus, for every word in a document from the Cantemist-NER corpus, its associated IOB2 label was assigned to all subtokens obtained from the same word.
2. **Multi-class fine-tuning.** Using the resulting Cantemist-NER corpus annotated with IOB2 tags at the subtoken-level, each transformer was fine-tuned on the automatic detection of tumor morphology mentions task. To perform the supervised fine-tuning of the whole model architecture on a multi-class sequence-labeling problem, the output representation encoded by the model for each subtoken was fed into a final fully-connected layer with 3 softmax

units, representing the “I”, “O” and “B” tags, respectively, of the IOB2 scheme.

3. **Subtoken-level predictions.** Hence, at inference time, given an input sequence of subwords as input to the model, the 3-tuple predicted for each subtoken could be interpreted as the probability of the subtoken being part of a word “inside” a tumor morphology mention (the “I” label), the probability of the subword belonging to a word “outside” any tumor morphology mention (the “O” label), and the probability of the subtoken being part of the “beginning” word of a tumor morphology mention (the “B” tag), respectively.
4. **Word-level predictions.** From the previous step, a set of IOB2 labels probabilities predicted by the model at the subtoken-level were obtained. However, in order to evaluate the predictive performance of the models on the Cantemist-NER task, the models predictions had to be converted into BRAT standoff format. Consequently, with the goal of transforming the IOB2 labels probabilities into BRAT format, we firstly converted the predictions made on the subtoken-level into word-level predictions. For this purpose, we applied a maximum probability criterion to the predictions made on the sequence of subtokens generated from each word. In this way, for the predictions made for all subtokens obtained from a single word, the criterion consisted in selecting, for each of the 3 IOB2 labels, the maximum predicted probability across the corresponding subtokens.
5. **Word-level tags.** Subsequently, considering the word-level predictions obtained from the previous step, each word was assigned the IOB2 tag predicted with the maximum probability. Then, using the IOB2 label associated to each word, the predictions made by the model were converted into BRAT format in order to evaluate the performance of the transformer on the Cantemist-NER task.

### 2.5 Experiments

We implemented our TL approach for tumor morphology mentions detection in TensorFlow, using the transformers library developed by HuggingFace [26]. For all transformer-based models analyzed in this study, we set a maximum input sequence length of 128 subwords. However, the majority of the clinical cases from the Cantemist-NER corpus have a subtoken sequence length clearly above 128 subwords. This represents a significant constraint when fine-tuning transformers on the Cantemist-NER task, since, for most of the documents, their whole sequence of subwords could not be used as input to the model. To overcome this limitation, we have used the fragment-based segmentation approach developed in [14]. In this way, each document from the Cantemist-NER corpus was firstly split into sentences. Then, adjacent sentences were grouped together in single fragments of text following a greedy strategy, in such a way that the subtokens sequence length of each fragment did not surpass the maximum input sequence length supported by the models. Finally, in order to fine-tune the transformers on the Cantemist-NER subtask, each generated text fragment was annotated with IOB2 labels at the subtoken-level, as described in the previous section. On the other hand, in the case of the feature-based CRF model, since no input sequence length limitation is imposed by this method, for every document in the Cantemist-NER corpus, its whole sequence of words annotated following the IOB2 tagging scheme was used to train the model. We used the sklearn-crfsuite[3] library to implement the CRF model, using traditional text mining features extracted from each word as input to the model, such as suffixes of 2 and 3 characters, boolean features indicating, for example, whether the word corresponds to a digit, and several features extracting information from nearby words. Finally, regarding the hardware resources employed, all experiments were conducted using a single GeForce GTX 1080 Ti GPU.

## 3 Results

Table 1 shows the performance of the 3 transformers on the Cantemist-NER task, as well as the performance of the baseline feature-based CRF model. For each transformer-based model, we compared the original general-domain version with the domain-specific version of the model adapted to the particularities of the Spanish oncology domain (see Section 2.4). The official evaluation metrics of the Cantemist-NER task [15]—precision, recall and F1-score—were employed to evaluate the predictive performance of the models. For each transformer, we fine-tuned 5 distinct randomly initialized instances. Comparing the performance of the baseline model with the results obtained by the transformer-based models, each transformer analyzed in this study significantly outperformed the feature-based CRF model for each of the three classification metrics described in Table 1. Among all models, mBERT-Galén, BETO-Galén and XLM-R-Galén achieved the best performance according to each classification metric, with the two domain-specific multilingual transformers—mBERT-Galén and XLM-R-Galén—obtaining identical average values for each metric, namely a mean precision of 0.867, an average recall of 0.869 and a mean F1-score of 0.868. On its part, the BETO-Galén model also obtained the same average F1-score of 0.868, but a slightly lower average recall (0.865) and a slightly greater average value for precision (0.872). Compared with the general-domain transformers, the domain-specific version of the models improved the performance for the detection of tumor morphology mentions in clinical reports in Spanish. In this way, for each transformer-based model, the clinical-domain version of the model outperformed the general-domain version in terms of the average values obtained for each classification metric. Finally, when comparing the results obtained in this study with the previously reported SOTA results, new SOTA performance was achieved according to the maximum values of each metric. Thus, the XLM-R-Galén model obtained a maximum precision value of 0.881, as well as a maximum recall of 0.878, exceeding the prior SOTA performance reported by the organizers of the Cantemist-NER task for each of the former two metrics [15]. In the

[3] https://sklearn-crfsuite.readthedocs.io/

case of the F1-score, the mBERT-Galén model surpassed the previous SOTA performance, obtaining a maximum value of 0.876.

**Table 1.** Models performances on the Cantemist-NER test set. The distribution of the precision, recall and F1-score values obtained by the 5 distinct fine-tuned instances of each model is described, by reporting the mean, standard deviation and maximum values. For the maximum values column of every metric, the best obtained result is bolded, while the second best is underlined.

| **Model** | **Precision** | | **Recall** | | **F1-score** | |
|---|---|---|---|---|---|---|
| | **Mean ± Std** | **Max** | **Mean ± Std** | **Max** | **Mean ± Std** | **Max** |
| Baseline-CRF | - | .815 | - | .774 | - | .794 |
| mBERT | .85 ± .009 | .861 | .854 ± .007 | .862 | .852 ± .004 | .858 |
| mBERT-Galén | .867 ± .008 | .876 | .869 ± .007 | .877 | .868 ± .004 | **.876** |
| BETO | .85 ± .006 | .859 | .858 ± .008 | .869 | .854 ± .004 | .856 |
| BETO-Galén | .872 ± .008 | .88 | .865 ± .004 | .869 | .868 ± .002 | .87 |
| XLM-R | .846 ± .014 | .861 | .858 ± .006 | .863 | .852 ± .005 | .858 |
| XLM-R-Galén | .867 ± .009 | **.881** | .869 ± .006 | **.878** | .868 ± .003 | .874 |
| Prior SOTA | - | .871 | - | .871 | - | .87 |

### 3.1 Ensemble

Additionally, we proposed an ensemble approach to combine the different IOB2 labels predictions made by the models at word-level. Hence, given a sequence of $W$ words, as a result of fine-tuning 5 different instances of each model, the fourth stage of our proposed workflow for performing tumor morphology mentions detection outputted 5 distinct IOB2 labels probability matrices of $W \times 3$ dimension (see Fig. 2) for a single transformer model. To merge these matrices into a single probability matrix, the proposed ensemble strategy consisted in performing the element-wise product of the 5 different matrices. Furthermore, our ensemble approach could also be employed to merge the IOB2 labels predictions made by any number of different transformers, by plainly performing the element-wise multiplication of all word-level IOB2 labels probability matrices obtained from the distinct models.

Table 2 describes the performance of our developed ensemble approach applied to merge both the word-level probabilities predicted by single models as well as the word-level predictions made by multiple distinct transformers. The ensemble combining the word-level predictions of the 3 transformer-based models adapted to the Spanish oncology domain—mBERT-Galén + BETO-Galén + XLM-R-Galén—achieved the best performance among all models examined in this work, obtaining a precision value of 0.893, a recall of 0.887 and a F1-score of 0.89, which remarkably surpassed the prior SOTA performance according to each classification metric.

**Table 2.** Ensemble models performances on the Cantemist-NER test subset, according to the precision, recall and F1-score metrics. For each metric, the best obtained result is bolded, while the second best is underlined.

| Ensemble | Precision | Recall | F1-score |
|---|---|---|---|
| mBERT | .873 | .872 | .872 |
| mBERT-Galén | .885 | .881 | .883 |
| BETO | .876 | .873 | .875 |
| BETO-Galén | .883 | .873 | .878 |
| XLM-R | .868 | .874 | .871 |
| XLM-R-Galén | .887 | .879 | .883 |
| mBERT + mBERT-Galén | .881 | .876 | .879 |
| BETO + BETO-Galén | .887 | .878 | .882 |
| XLM-R + XLM-R-Galén | .883 | .88 | .882 |
| mBERT + BETO + XLM-R | .882 | .876 | .879 |
| mBERT-Galén + BETO-Galén + XLM-R-Galén | **.893** | **.887** | **.89** |
| Prior SOTA | .871 | .871 | .87 |

## 4 Conclusion

In this work, we systematically examined the performance of 3 transformer-based models to perform the detection of tumor morphology mentions in clinical documents in Spanish. Using a TL-based strategy, the transformers were first adapted to the particularities of the Spanish oncology domain by pretraining the models on a real-world corpus of oncology clinical cases written in Spanish. Subsequently, the resulting models were fine-tuned on the Cantemist-NER corpus, following a multi-class sequence-labeling approach. For each analyzed transformer, the domain-specific version outperformed the general-domain version of the model on the Cantemist-NER task. Finally, the combination of the TL-based approach with an ensemble strategy that exploited the predictive capacities of the 3 different transformers, yielded the best achieved results, which noticeably improved the prior SOTA performance on the Cantemist-NER task. In future works, given the promising results obtained in this paper, we will try to extend the TL-based methodology to perform other downstream medical NLP tasks in Spanish using transformers, such as the de-identification of a real-world clinical corpus or the NER of cancer prognostic factors in medical records.

**Acknowledgments.** This work was partially supported by the project PID2020-116898RB-I00, Ministerio de Ciencia e Innovación, Plan Nacional de I+D+i, the project UMA-CEIATECH-01, Andalucía TECH, and the I Plan Propio de Investigación, Transferencia y Divulgación Científica of the Universidad de Málaga.

## References

1. Alsentzer, E., Murphy, J., Boag, W., Weng, W.H., Jindi, D., Naumann, T., McDermott, M.: Publicly available clinical BERT embeddings. In: Proceedings of the 2nd Clinical Natural Language Processing Workshop. pp. 72–78. Association for Computational Linguistics, Minneapolis, Minnesota, USA (Jun 2019)
2. Baumann, L.A., Baker, J., Elshaug, A.G.: The impact of electronic health record systems on clinical documentation times: A systematic review. Health Policy **122**(8), 827–836 (Aug 2018)
3. Bronnert, J.: Preparing for the CAC transition. J. AHIMA **82**(7), 60–1; quiz 62 (Jul 2011)
4. Cañete, J., Chaperon, G., Fuentes, R., Ho, J.H., Kang, H., Pérez, J.: Spanish pre-trained BERT model and evaluation data. In: Practical ML for Developing Countries Workshop@ ICLR 2020 (2020)
5. Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., Stoyanov, V.: Unsupervised cross-lingual representation learning at scale. arXiv [cs.CL] (Nov 2019)
6. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of deep bidirectional transformers for language understanding. arXiv [cs.CL] (Oct 2018)
7. García-Pablos, A., Perez, N., Cuadros, M.: Vicomtech at CANTEMIST 2020. In: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2020). pp. 489–498. CEUR Workshop Proceedings (2020)
8. Hughes, M., Li, I., Kotoulas, S., Suzumura, T.: Medical text classification using convolutional neural networks. Stud. Health Technol. Inform. **235**, 246–250 (2017)
9. Johnson, M., Schuster, M., Le, Q.V., Krikun, M., Wu, Y., Chen, Z., Thorat, N., Viégas, F., Wattenberg, M., Corrado, G., Hughes, M., Dean, J.: Google's multilingual neural machine translation system: Enabling Zero-Shot translation. Transactions of the Association for Computational Linguistics **5**, 339–351 (2017)
10. Kudo, T., Richardson, J.: SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. arXiv [cs.CL] (Aug 2018)
11. Lafferty, J.D., McCallum, A., Pereira, F.C.N.: Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In: Proceedings of the Eighteenth International Conference on Machine Learning. pp. 282–289. ICML '01, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA (Jun 2001)
12. Lample, G., Conneau, A.: Cross-lingual language model pretraining. arXiv [cs.CL] (2019)
13. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V.: RoBERTa: A robustly optimized BERT pretraining approach. arXiv [cs.CL] (2019)
14. López-García, G., Jerez, J.M., Ribelles, N., Alba, E., Veredas, F.J.: ICB-UMA at CANTEMIST 2020: Automatic ICD-O coding in spanish with BERT. In: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2020). pp. 468–476. CEUR Workshop Proceedings (2020)

15. Miranda-Escalada, A., Farré-Maduell, E., Krallinger, M.: Named Entity Recognition, Concept Normalization and Clinical Coding: Overview of the Cantemist Track for Cancer Text Mining in Spanish, Corpus, Guidelines, Methods and Results. In: Iberian Languages Evaluation Forum (IberLEF 2020). pp. 303–323. CEUR Workshop Proceedings, Málaga, Spain (2020)
16. Mujtaba, G., Shuib, L., Idris, N., Hoo, W.L., Raj, R.G., Khowaja, K., Shaikh, K., Nweke, H.F.: Clinical text classification research trends: Systematic literature review and open issues. Expert Syst. Appl. **116**, 494–520 (Feb 2019)
17. National Cancer Institute: How Cancer Is Diagnosed. https://www.cancer.gov/about-cancer/diagnosis-staging/diagnosis (2019), Accessed: 2021-04-23
18. Qiu, J.X., Yoon, H.J., Fearn, P.A., Tourassi, G.D.: Deep Learning for Automated Extraction of Primary Sites From Cancer Pathology Reports. IEEE Journal of Biomedical and Health Informatics **22**(1), 244–251 (2018)
19. Ramshaw, L.A., Marcus, M.P.: Text chunking using Transformation-Based learning. In: Armstrong, S., Church, K., Isabelle, P., Manzi, S., Tzoukermann, E., Yarowsky, D. (eds.) Natural Language Processing Using Very Large Corpora, pp. 157–176. Springer Netherlands, Dordrecht (1999)
20. Ribelles, N., Jerez, J.M., Urda, D., Subirats, J.L., Márquez, A., Quero, C., Franco, L.A.: Galén: Sistema de información para la gestión y coordinación de procesos en un servicio de oncología. RevistaeSalud **6**(21) (2010)
21. Si, Y., Wang, J., Xu, H., Roberts, K.: Enhancing clinical concept extraction with contextual embeddings. J. Am. Med. Inform. Assoc. **26**(11), 1297–1304 (Nov 2019)
22. Stenetorp, P., Pyysalo, S., Topić, G., Ohta, T., Ananiadou, S., Tsujii, J.: BRAT: a Web-based Tool for NLP-Assisted Text Annotation. In: Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics. pp. 102–107. Association for Computational Linguistics, Avignon, France (Apr 2012)
23. Sung, H., Ferlay, J., Siegel, R.L., Laversanne, M., Soerjomataram, I., Jemal, A., Bray, F.: Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin (2021)
24. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L.U., Polosukhin, I.: Attention is all you need. In: Advances in Neural Information Processing Systems 30, pp. 5998–6008. Curran Associates, Inc. (2017)
25. Vítores, D.F.: El español: una lengua viva. Informe 2020. Instituto Cervantes (2020)
26. Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Le Scao, T., Gugger, S., Drame, M., Lhoest, Q., Rush, A.M.: HuggingFace's transformers: State-of-the-art natural language processing. arXiv [cs.CL] (Oct 2019)
27. Xiong, Y., Huang, Y., Chen, Q., Wang, X., Nic, Y., Tang, B.: A Joint Model for Medical Named Entity Recognition and Normalization. In: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2020). pp. 499–504. CEUR Workshop Proceedings (2020)
28. Yang, X., Bian, J., Hogan, W.R., Wu, Y.: Clinical concept extraction using transformers. J. Am. Med. Inform. Assoc. **27**(12), 1935–1942 (Dec 2020)
29. Zhu, F., Patumcharoenpol, P., Zhang, C., Yang, Y., Chan, J., Meechai, A., Vongsangnak, W., Shen, B.: Biomedical text mining and its applications in cancer research. Journal of Biomedical Informatics **46**(2), 200–211 (2013)